Section 1.3

Check Your Understanding, page 53:

1. Because the distribution is skewed to the right, we would expect the mean to be larger than the median.

2. The mean is $\left(\dfrac{5+10+10+15+...+85}{20}\right)=31.25$ minutes, which is bigger than the median of 22.5 minutes.

3. Because the distribution is skewed, the median would be a better measure of the center of the distribution.

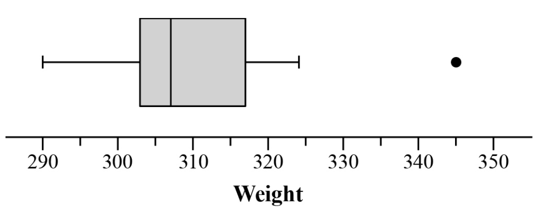
Check Your Understanding, page 59:

1. The data in order are: 290, 301, 305, 307, 307, 310, 324, 345. The minimum is 290 and the maximum is 345. The median is the average of the 4th and 5th observations: median = $\left(\dfrac{307+307}{2}\right)=307$. The first quartile is the median of the 4 values below the median: Q_1 = $\left(\dfrac{301+305}{2}\right)=303$. The third quartile is the median of the 4 values above the median: Q_3 = $\left(\dfrac{310+324}{2}\right)=317$. The five-number summary is 290, 303, 307, 317, 345.

2. The *IQR* is 317 – 303 = 14 pounds. The range of the middle half of the data is 14 pounds.

3. 1.5*IQR* = 1.5(14) = 21. Any outliers occur below 303 – 21 = 282 or above 317 + 21 = 338. There are no observations below 282. However, there is one observation above 338. The value of 345 pounds is an outlier.

4. The boxplot is given below.



Check Your Understanding, page 63:

1. The mean is $\left(\dfrac{67+72+76+76+84}{5}\right)=75$.

2. The table is given below:

Observation	Deviation	Squared Deviation
67	$67-75=-8$	$(-8)^2=64$
72	$72-75=-3$	$(-3)^2=9$
76	$76-75=1$	$1^2=1$
76	$76-75=1$	$1^2=1$
84	$84-75=9$	$9^2=81$
Total	0	156

3. The variance is the sum of the squared deviations divided by 4 (one less than the number of observations). The variance is $\frac{156}{4} = 39$ inches squared. The standard deviation is the square root of the variance, $s_x = \sqrt{39} = 6.24$ inches.
4. The players' heights typically vary by about 6.24 inches from the mean height of 75 inches.

Exercises, page 69:

1.79 The mean of Joey's first 14 quiz scores is $\frac{86+84+...+93}{14} = \frac{1190}{14} = 85$.

1.80 The mean weight for the 7 defensive linemen on the 2009 Dallas Cowboys is
$\frac{321+285+300+285+286+293+298}{7} = 295.4$ pounds.

1.81 (a) Putting the scores in order: 74, 75, 76, 78, 80, 82, 84, 86, 87, 90, 91, 93, 96, 98. Because there are 14 scores, the median is the average of the 7th and 8th scores. Therefore the median is $\frac{84+86}{2} = 85$.
(b) If Joey had a 0 for the 15th quiz then the sum of his quiz scores would still be 1190, leading to a mean of $\frac{1190}{15} = 79.33$. To find the median, we add the 0 to the beginning of the list in part (a). Since there are now 15 measurements, the median would be the 8th measurement, which is 84. Notice that the median did not change much but the mean did. This shows that the mean is not resistant to outliers, but the median is.

1.82 (a) Putting the weights in order: 285, 285, 286, 293, 298, 300, 321. Because there are 7 values, the median is the 4th value. Therefore the median is 293 pounds.
(b) If the heaviest weight were 341 instead of 321, then the mean would increase to
$\frac{285+285+286+293+298+300+341}{7} = 298.3$, but the median would not change. This is because the median is resistant to outliers, but the mean is not.

1.83 The mean is \$60,954 and the median is \$48,097. The distribution of salaries is likely to be quite right skewed because of a few people who have a very large income. When a distribution is skewed to the right, the mean is bigger than the median.

1.84 The mean house price is \$263,200 and the median is \$224,200. The distribution of house prices is likely to be quite skewed to the right because of a few very expensive homes. When a distribution is skewed to the right, the mean is bigger than the median.

1.85 The team's annual payroll is $1.2(25) = 30$ or \$30 million. No, you would not be able to calculate the team's annual payroll from the median because the median only describes the middle value in the distribution. It doesn't provide specific information about any of the other values.

1.86 The mean salary is $\dfrac{5(22,000)+2(50,000)+270,000}{8}$ = \$60,000. Seven of the eight employees (everyone but the owner) earned less than the mean. The median is \$22,000. An unethical recruiter would report the mean salary as the "typical" or "average" salary. However, the median is a more accurate depiction of a "typical" employee's earnings, because it is not influenced by the outlier of \$270,000.

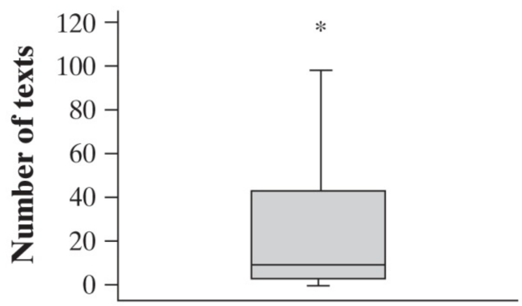
1.87 (a) Estimate the frequencies of the bars (from left to right): 10, 40, 42, 58, 105, 60, 58, 38, 27, 18, 20, 10, 5,5,1 and 3 (although answers may vary slightly, the frequencies must sum to 500). Using these values, we can estimate the mean by adding 2 ten times, 3 forty times, and so on. This gives us a sum of 3504. The mean is then calculated by dividing by the number of responses: $\bar{x} = \dfrac{3504}{500} = 7.01$. Alternatively, we can estimate the mean by finding the balance point of the distribution, which is approximately 7. We estimate the median by finding the average of the 250th and 251st values. The median is 6.
(b) Because the median is less than the mean (which is not surprising in a right skewed distribution), we would use the median to argue that shorter domain names are more popular.

1.88 (a) Estimate the frequencies of the bars (from left to right): 15, 11, 15, 11, 8, 5, 3, 3, 3 (although answers may vary slightly, the frequencies must sum to 74). We estimate the median by finding the average of the 37th and 38th values. The median is 2. The first quartile is the median of the 37 observations below the median, which is the 19th observation. Thus, $Q_1 = 1$. The third quartile is the median of the 37 observations above the median, which is the 56th observation. Thus, $Q_3 = 4$.
(b) Using these values, we can estimate the mean by adding 0 fifteen times, 1 eleven times, and so on. This gives us a sum of 194. The mean is then calculated by dividing by the number of responses: $\bar{x} = \dfrac{194}{74} = 2.62$. Alternatively, we can estimate the mean by finding the balance point of the distribution , which appears to be a little higher than 2.5.

1.89 (a) Putting the scores in order: 74, 75, 76, 78, 80, 82, 84, 86, 87, 90, 91, 93, 96, 98. There are 14 observations, so the first quartile is the median of the 7 observations below the median (85). This means that it is the value of the 4th observation which is $Q_1 = 78$ points. The third quartile is the median of the 7 observations above the median (85), so it is the value of the 11th observation, which is $Q_3 = 91$ points. So, $IQR = 91 - 78 = 13$. The middle 50% of the data have a range of 13 points.
(b) Any outliers are below $Q_1 - 1.5\text{IQR}$ or above $Q_3 + 1.5\text{IQR}$. These boundaries are $78 - 1.5(13) = 58.5$ and $91 + 1.5(13) = 110.5$. There are no points outside of these boundaries, so there are no outliers.

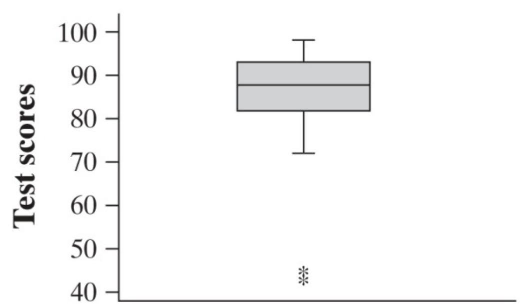
1.90 (a) Putting the weights in order: 285, 285, 286, 293, 298, 300, 321. There are 7 observations, so the first quartile is the median of the 3 observations below the median (293). This means that it is the value of the 2nd observation, which is $Q_1 = 285$ pounds. The third quartile is the median of the 3 observations above the median (293), so it is the value of the 6th observation, which is $Q_3 = 300$ pounds. So, $IQR = 300 - 285 = 15$ pounds. The middle 50% of the data have a range of 15 pounds.
(b) Any outliers are below $Q_1 - 1.5\text{IQR}$ or above $Q_3 + 1.5\text{IQR}$. These boundaries are 285 – 1.5(15) = 262.5 and 300 + 1.5(15) = 322.5. There are no points outside of these boundaries, so there are no outliers.

1.91 (a) Putting the data in order: 0, 0, 0, 1, 1, 3, 3, 5, 5, 7, 8, 8, 9, 14, 25, 25, 26, 29, 42, 44, 52, 72, 92, 98, 118. The median is 9, the first quartile is 3, and the third quartile is 43. The *IQR* is 43 – 3 = 40. Outliers are anything below $3-1.5(40)=-57$ or above $43+1.5(40)=103$. This means that the value of 118 is an outlier. The boxplot is shown below.



(b) The article claims that teens send 1742 texts a month, which works out to be about 58 texts a day (assuming a 30 day month). Nearly all of the members of the class (21 of 25) sent fewer than 58 texts per day, which seems to contradict the claim in the article.

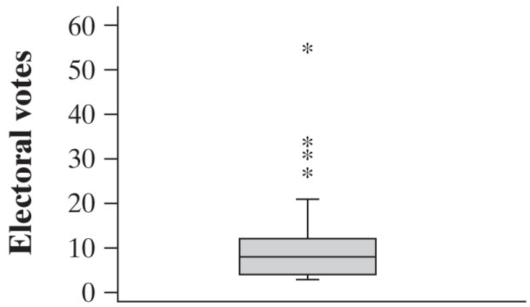
1.92 (a) Putting the data in order: 43, 45, 72, 73, 78, 80, 81, 82, 82.5, 85, 85.5, 86, 86, 87.5, 87.5, 88, 88, 89, 91, 91, 92, 93, 93, 93.5, 93.5, 94.5, 94.5, 95, 96, 98. The median is 87.75, the first quartile is 82, and the third quartile is 93. The *IQR* is 93 – 82 = 11. Any observation above $Q_3+1.5IQR=93+1.5(11)=109.5$ or below $Q_1-1.5IQR=82-1.5(11)=65.5$ is considered an outlier. Thus, the scores 43 and 45 are outliers. The boxplot is shown below.



(b) Most students did quite well. In fact, more than 75% of the class scored higher than 80. Only two of the students did very poorly.

1.93 (a) Since the data are recorded as texts – calls, positive numbers indicate students who had more text messages than calls. It appears from the boxplot as though the first quartile is 0 which means that 75% of the students had more texts than calls. This does support the article's conclusion.
(b) No, we cannot make any more general conclusion. The students in a statistics class might not be representative of all students at the school. For example, students in statistics classes tend to be upperclassmen and their responses might differ from underclassmen.

1.94 (a) The median is 8, the first quartile is 4, and the third quartile is 12. The *IQR* is 12 – 4 = 8. An outlier is any observation that is less than $Q_1 - 1.5IQR = 4 - 1.5(8) = -8$ or greater than $Q_3 + 1.5IQR = 12 + 1.5(8) = 24$. Thus, there are 4 outliers: 27, 31, 34 and 55. The boxplot is shown below.



(b) Use the median and *IQR* rather than the mean and standard deviation because the distribution is skewed with several high outliers.

1.95 (a) The stock fund varied between about -3.5% and 3%.
(b) The median return was about 0.1% for both funds.
(c) The stock fund is much more variable. It has higher positive returns, but also higher negative returns.

1.96 All five income distributions are skewed to the right. As education level rises, the median, quartiles, and extremes rise as well—that is, all five points on the boxplot get larger. Additionally, the variability in income increases as the education level increases. Both the width of the box (the *IQR*) and the distance from one extreme to the other increase as education levels increase.

1.97 (a) The mean phosphate level is $\bar{x} = \dfrac{32.4}{6} = 5.4$ mg/dl.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5.6	0.2	0.04
5.2	–0.2	0.04
4.6	–0.8	0.64
4.9	–0.5	0.25
5.7	0.3	0.09
6.4	1.0	1.00
32.4	0	2.06

The variance is $s_x^2 = \dfrac{2.06}{5} = 0.412$ and the standard deviation is $s_x = \sqrt{\dfrac{2.06}{5}} = 0.6419$ mg/dl.
(b) The phosphate level typically varies from the mean by about 0.6419 mg/dl.

1.98 (a) Mean = $\bar{x} = \dfrac{7+7+9+9}{4}$ = 8 hours.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	–1	1
7	–1	1
9	1	1
9	1	1
32	0	4

The variance is $s_x^2 = \dfrac{4}{3} = 1.33$ and the standard deviation is $s_x = \sqrt{\dfrac{4}{3}}$ = 1.15 hours.
(b) The hours of sleep typically varies from the mean by about 1.15 hours.
(c) No, it would not be safe to make this generalization. The first 4 students to arrive in the classroom are not likely to be representative of the entire class in terms of the amount of sleep they got last night.

1.99 (a) It looks like the distribution is skewed to the right because the mean is much larger than the median. Also, Q_3 is much further from the median than Q_1.
(b) The amount of money spent typically varies from the mean by $21.70.
(c) The first quartile is 19.06 and the third quartile is 45.72 so the *IQR* is 45.72 – 19.06 = 26.66. Any points below 19.06 – 1.5(26.66) = –20.93 or above 45.72 + 1.5(26.66) = 85.71 are outliers. Because the maximum of 93.34 is greater than 85.71, there is at least one outlier.

1.100 (a) It would appear that the distribution for the female doctors is more likely to be symmetric since the mean and median are relatively close together (19.1 and 18.5 respectively). The mean and median for the male doctors are quite far apart (41.333 and 34 respectively).
(b) The *IQR* measures the range of the middle 50% of the data and does not consider the lower 25% or the upper 25% of the data. The standard deviation, however, uses every observation and is not resistant to outliers. So, while the middle 50% of the data sets may have similar spreads, if the lower and upper 25% of one distribution are more spread out than the other distribution, it will have a larger standard deviation.
(c) It does appear that males perform more C-sections. Each of the numbers in the 5-number summary was larger for the males than for the females.

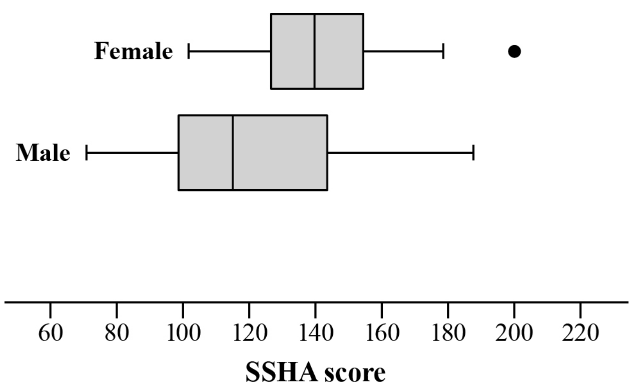
1.101 Yes, the *IQR* is a resistant measure of spread. For example, consider the simple data set 1, 2, 3, 4, 5, 6, 7, 8. The median = 4.5, $Q_1 = 2.5$, $Q_3 = 6.5$, and $IQR = 4$. Changing any value outside the interval between Q_1 and Q_3 will have no effect on the *IQR*. For example, if 8 is changed to 88, the *IQR* will still be 4.

1.102 (a) This could be used to measure the center since we are averaging the 25th and 75th percentiles, effectively finding a middle point between these positions. It would be resistant to outliers, because any outliers would occur below Q_1 or above Q_3.
(b) This could be used as a measure of spread since it finds the distance between the smallest and largest values and then divides by 2, giving half of the range. This measure, however, would not be resistant to outliers. If outliers exist, they would, by definition, include the max, the min, or both.

1.103 (a) One possible answer is 1, 1, 1, 1.
(b) 0, 0, 10, 10.
(c) For part (a), any set of four identical numbers will have $s_x = 0$. For part (b), however, there is only 1 possible answer. We want the values to be as far from the mean as possible, so the squared deviations from the mean can be as big as possible. If we choose 0, 10, 10, 10—or 10, 0, 0, 0—we make the first squared deviation 7.5^2, but the other three are only 2.5^2. Our best choice is two values at each extreme, which makes all four squared deviations equal to 5^2.

1.104 Variable A has a larger standard deviation because more of the observations have values further from the mean. Because of the bell-shape to the distribution of variable B, more of the observations have values quite close to the mean, making the typical distance from the mean smaller.

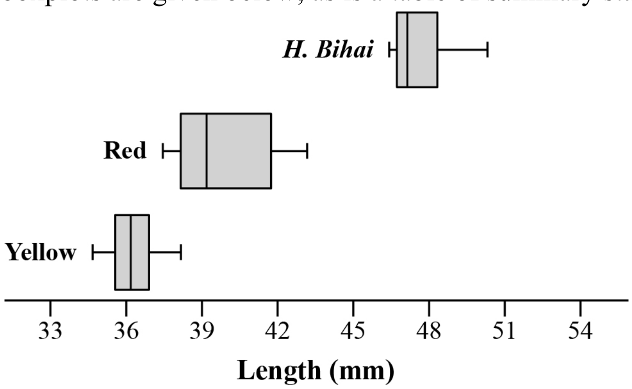
1.105 *State:* Do the data indicate that men and women differ in their study habits and attitudes towards learning? *Plan:* We will draw side-by-side boxplots of the data about men and women. Then we'll compute summary statistics and compare the shape, center, and spread of both distributions. *Do:* The boxplots are given below, as is a table of summary statistics.



Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Women	18	141.06	26.44	101.00	126.00	138.50	154.00	200.00
Men	20	121.25	32.85	70.00	98.00	114.50	143.00	187.00

Both distributions are slightly skewed to the right. Both the mean and median are higher for women (mean = 141.06, median = 138.5) than for men (mean = 121.25, median = 114.5). The scores for men are more variable (s_x = 32.85, *IQR* = 45) than the scores for women (s_x = 26.44, *IQR* = 28). There are no outliers in the male distribution, and a single outlier at 200 in the female distribution. *Conclude:* It appears from the boxplot and the numerical summaries that men and women differ in their study habits and attitudes towards learning. The typical score for females is about 24 greater than the typical score for males. Female scores are also more consistent than male scores.

1.106 *State:* Do the different types of flowers have different distributions of length? *Plan:* We will draw side-by-side boxplots of the data about each type of flower. Then we'll compute summary statistics and compare the shape, center, and spread of the distributions. *Do:* The boxplots are given below, as is a table of summary statistics.



Variable	N	Mean	StDev
H. bihai	16	47.597	1.213
red	23	39.711	1.799
yellow	25	36.180	0.975

Variable	Minimum	Q1	Median	Q3	Maximum
H. bihai	46.340	46.710	47.120	48.245	50.260
red	37.400	38.070	39.160	41.690	43.090
yellow	34.570	35.450	36.110	36.820	38.130

The distributions for *H. bihai* and *H. caribaea red* are skewed to the right while the distribution of *H. caribaea yellow* is roughly symmetric. Both the mean and median length are greatest for *H. bihai* (mean = 47.597, median = 47.12). Also, the mean and median for *H. caribaea red* are greater (mean = 39.711, median 39.16) than for *H. caribaea yellow* (mean = 36.180, median = 36.11). The lengths for *H. caribaea red* are the most variable (s_x = 1.799, *IQR* = 3.62) and *H. bihai* are the second most variable (*H. bihai*: s_x = 1.213, *IQR* = 1.535; *H. caribaea yellow*: s_x = 0.975, *IQR* = 1.37). There are no outliers in any of the distributions.
Conclude: It appears from the boxplots and summary statistics that the three flower varieties have distinct distributions of lengths. *H. bihai* is clearly the longest variety—the shortest *bihai* was over 3 mm longer than the tallest red. Red is generally longer than yellow, with a few exceptions.

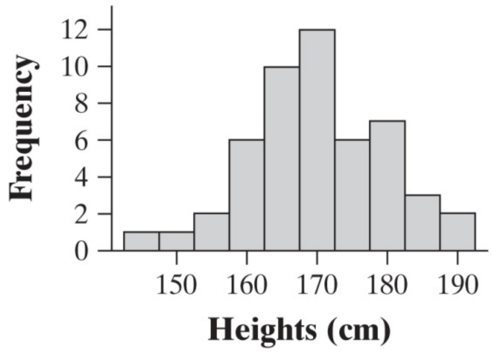
1.107 d

1.108 b

1.109 e

1.110 a

1.111 A histogram is given below:



This distribution is roughly symmetric with a center around 170 cm and values that vary from 145.5 cm to 191 cm. There do not appear to be any outliers.

1.112 (a) Yes, a pie chart is appropriate here because the categories (method of communication) form parts of a whole.
(b) The graph should not be described as skewed to the right because this is a distribution of categorical data, not quantitative data. The categories could be graphed in any order.

1.113 Women appear to be more likely to engage in behaviors that are indicative of good "habits of mind." They are especially more likely to revise papers to improve their writing (about 55% of females report this as opposed to about 37% of males). The difference is a little less for seeking feedback on their work. In that case about 49% of the females did this as opposed to about 38% of males.